Submitted 6/18; Revised 10/18; Published 11/18

DALEX: Explainers for Complex Predictive Models in R

Przemysław Biecek PRZEMYSLAW.BIECEK@GMAIL.COM
Faculty of Mathematics and Information Science, Warsaw University of Technology
75 Koszykowa Street, Warsaw, Poland
Samsung Research Poland

Editor: Alexandre Gramfort

Abstract

Predictive modeling is invaded by elastic, yet complex methods such as neural networks or ensembles (model stacking, boosting or bagging). Such methods are usually described by a large number of parameters or hyper parameters - a price that one needs to pay for elasticity. The very number of parameters makes models hard to understand.

This paper describes a consistent collection of explainers for predictive models, a.k.a. black boxes. Each explainer is a technique for exploration of a black box model. Presented approaches are model-agnostic, what means that they extract useful information from any predictive method irrespective of its internal structure. Each explainer is linked with a specific aspect of a model. Some are useful in decomposing predictions, some serve better in understanding performance, while others are useful in understanding importance and conditional responses of a particular variable.

Every explainer presented here works for a single model or for a collection of models. In the latter case, models can be compared against each other. Such comparison helps to find strengths and weaknesses of different models and gives additional tools for model validation. Presented explainers are implemented in the `DALEX` package for R. They are based on a uniform standardized grammar of model exploration which may be easily extended.

Keywords: interpretable machine learning, explainable artificial intelligence, predictive modelling, model visualization

1. Introduction

Predictive modeling has a large number of applications in almost every area of human activity, starting from medicine, marketing, logistic, banking and many others. Due to the increasing amount of collected data, models become more sophisticated and complex.

It is believed that there is a trade-off between the interpretability and accuracy of a model (see Johansson et al., 2011). It comes from the observation that the most elastic models usually have higher accuracy but in turn they are also more complex. Complexity here means a large number of model parameters that affect the final prediction. That number is big enough to make the model ununderstandable for an ordinary human being.

In many areas we cannot sacrifice interpretability, either because of legal requirements (see *right to explanation* in GDPR), or because it leads to unfair decisions (see O'Neil, 2016) or because it is important for users (see Lundberg and Lee, 2017). Interpretability brings multiple benefits such as: a) helps to extract interpretable patterns from trained models; b) helps to identify reasons behind poor predictions; c) increases trust in model predictions (see Ribeiro et al., 2016); d) reduces the hidden debt in machine learning models (see Sculley et al., 2015); e) helps to detect bias in machine learning models; f) creates additional safety catch that may protect from overfitted models.

In this paper we present a consistent general framework for exploration of black-box models. This framework covers the most known approaches to interpretability and structure exploration, such as Partial Dependence Plots (Greenwell, 2017), Accumulated Local Effects Plots (Apley, 2017), Merging Path Plots (Sitko and Biecek, 2017), Break Down Plots (Staniak and Biecek, 2018), Permutational Variable Importance Plots (Fisher et al., 2018) or Cateris Paribus Plots. An unique feature of `DALEX` explainers is that they can be natively used to compare two or more models. Model comparison helps to understand differences in model responses, gives new insights that may be used to construct new, better features.

Presented framework is available as an open source package `DALEX` for R. The R language (R Core Team, 2017) is one of the most popular languages for statistical and machine learning modeling. `DALEX` works with any predictive model. The extended user documentation[1] contains examples for the most popular frameworks, such as `caret` (Kuhn, 2008), `mlr` (Bischl et al., 2016), Random Forest and Gradient Boosting Machines. The `DALEX` package is available on at CRAN and GitHub[2] along with technical documentation[3].

Example explainers presented in this paper were recorded with the `archivist` package (Biecek and Kosinski, 2017). To save space, we present only graphical explainers. Numerical explainers can be downloaded with R commands listed in footnotes.

2. Architecture

Figure 1 presents the general architecture of the `DALEX` package. This methodology is model-agnostic and works for predictive models, such as classification or regression models.

Methods for understanding of global structure of a model (a.k.a. model explainers) and for understanding of a local structure of a model (a.k.a. prediction explainers) are implemented in separate functions. We call these functions *explainers* since they are designed to explain a single feature of a model. Every explainer returns numerical summaries in a tabular format. These tables may be visualized with generic `plot` function. The `plot` function works also for multiple models and overlays model explainers in a single chart. See examples in Figure 1 panels F, I and J.

2.1. Prediction Understanding: Explainers for Variable Attribution

The most known approaches to explanations of a single prediction are *LIME* method (Ribeiro et al., 2016), for local variable importance, and *Shapley values* (Lundberg and Lee, 2017), for local variable attribution. *Break Down Plots* are fast approximations of *Shapley values*. Comparison of these methods is presented in Staniak and Biecek (2018). An example for these explainers[4] is presented in Figure 1 panels C and D.

Note, that for non additive models, the local model behaviour may be very different from global model behaviour. Consider $f(x_1, x_2) = x_1 * x_2$ around point $(0, 0)$.

[1] User documentation is available at `https://pbiecek.github.io/DALEX_docs`.

[2] Development version is available at `https://github.com/pbiecek/DALEX`.

[3] Technical documentation is available at `https://pbiecek.github.io/DALEX`.

[4] Access this explainer with `archivist::aread('pbiecek/DALEX_arepo/72b47')`.

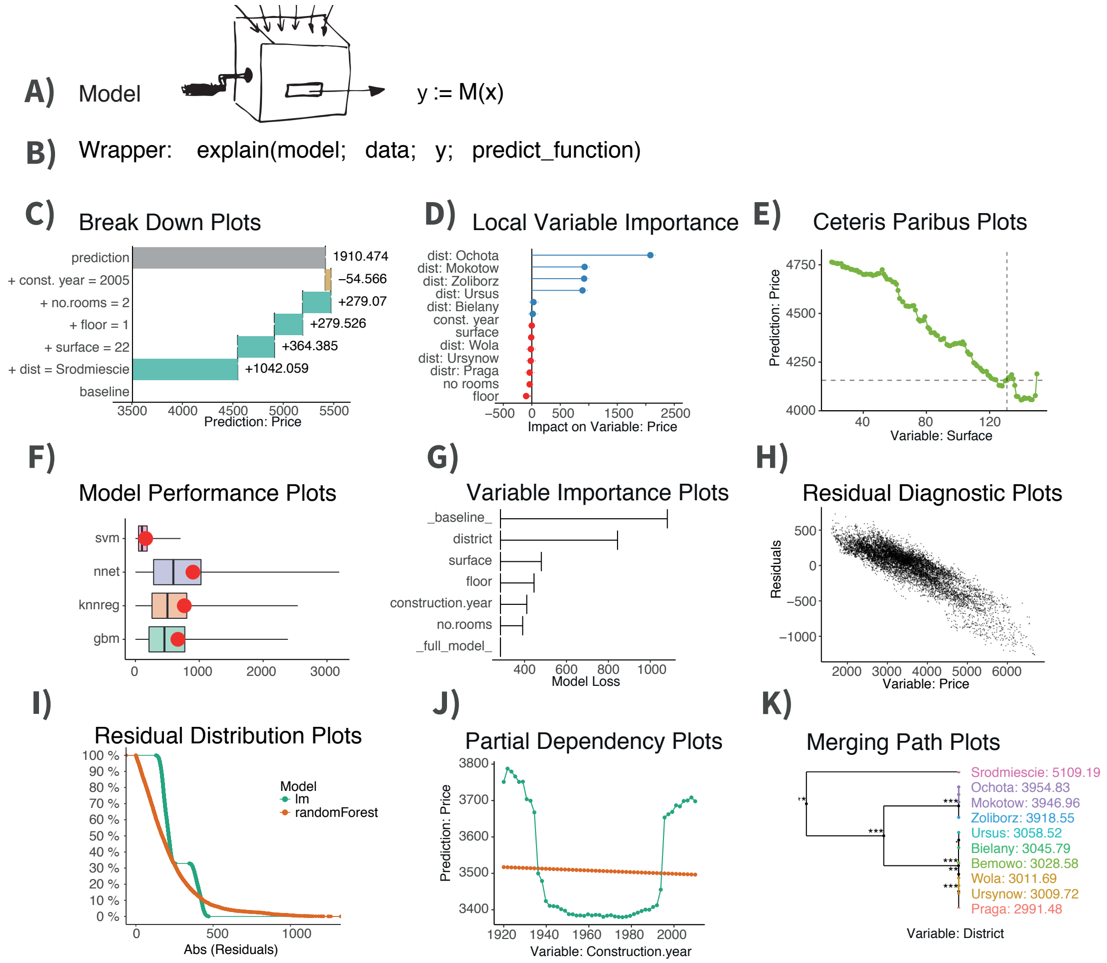

Figure 1: Architecture of the DALEX package is based on simple unified grammar. A) Any predictive model $M : \mathcal{R}^p \to \mathcal{R}$ may be used. B) Models are first enriched with additional metadata with the function `explain()`. Each explainer returns numerical summaries that can be plotted with generic `plot()` function. C, D, E) Explainers for a single prediction. F) Comparison of four models. G, H) Explainers for a single model, I) Comparison of residuals for two models. J, K) Explainers for a single variable, respectively continuous and categorical.

2.2. Prediction Understanding: Explainers for What-If Scenarios

Ceteris Paribus Profiles show how the model response changes as a function of a single variable. These plots recollect similarities to more known *Partial Dependency Plots*. The difference between them is that *Ceteris Paribus Profiles* are focused on a single observation.

An example for this explainer[5], is presented in Figure 1 panel E. One can read how model response will change for an altered value of a single variable.

[5] Access this explainer with `archivist::aread('pbiecek/DALEX_arepo/c8989')`.

2.3. Model Understanding: Explainers for Model Performance

Model performance is often summarized with a single number such as *F1* or *accuracy.* This makes it easier to construct a ranking of models and choose the best one. However, more descriptive statistics are better when it comes to understanding of a model. The descriptive statistics most often used for classification is *ROC (Receiver Operating Characteristic)* with various extensions for regression as in Hernndez-Orallo (2013).

The DALEX package offers a selection of tools for exploration of model performance, see Figure 1 panels F and I[6], and model diagnostic, see Figure 1 panel H. The latter is available through the auditor package (Gosiewska and Biecek, 2018), closely integrated with DALEX.

2.4. Model Understanding: Explainers for Effect of a Single Variable

The DALEX package offers a selection of tools for better understanding of a conditional model's response based on a single variable. For continues variables it supports *Partial Dependence Plot* (Greenwell, 2017) as implemented in the pdp package and *Accumulated Local Effects Plot* (Apley, 2017) as implemented in ALEPlot package, see Figure 1 panel J[7]. For categorical variables it supports *Merging Path Plot* (Sitko and Biecek, 2017) as implemented in the factorMerger package. See Figure 1 panel K.

2.5. Model Understanding: Explainers for Variable Importance

The DALEX package offers a model-agnostic procedure to calculate variable importance. The model-agnostic approach is based on permutational approach introduced initially for *Random Forest* (Breiman, 2001) and then extended for other models by Fisher et al. (2018).

An example for these explainers[8] is presented in Figure 1 panel G. It's common in variable importance charts to hitch bars in 0. Charts in the DALEX package present not only drop in model performance but also the initial model performance. In that way one can compare variables between models with different initial performance.

3. Summary

In this article we have introduced consistent methodology and tools for model-agnostic explanations. Global explainers (for model understanding) and local explainers (for prediction understanding) are based on uniform grammar. Every explainer creates a numerical summary, visual summary and allows for comparison of multiple models. The DALEX package is tested with CI tools and is easy to extend[9]. Here we presented DALEX 0.2.5 with R 3.5.1.

Acknowledgments

The work was financially supported by RENOIR Project under the Marie Skłodowska-Curie grant agreement No 691152 and NCN Opus grant 2016/21/B/ST6/02176.

[6] Access this explainer with archivist::aread('pbiecek/DALEX_arepo/b4eb1').

[7] Access this explainer with archivist::aread('pbiecek/DALEX_arepo/3b150').

[8] Access this explainer with archivist::aread('pbiecek/DALEX_arepo/9378c').

[9] Extended version of this paper is available at https://arxiv.org/pdf/1806.08915v1.pdf.

References

Dan Apley. *ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots*, 2017. R package version 1.0.

Przemyslaw Biecek and Marcin Kosinski. archivist: An R Package for Managing, Recording and Restoring Data Analysis Results. *Journal of Statistical Software*, 82(11):1–28, 2017. doi: 10.18637/jss.v082.i11.

Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. mlr: Machine Learning in R. *Journal of Machine Learning Research*, 17(170):1–5, 2016.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. *Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective*, 2018. URL `http://arxiv.org/abs/1801.01489`.

Alicja Gosiewska and Przemyslaw Biecek. *auditor: an R Package for Model-Agnostic Visual Validation and Diagnostic*, 2018. URL `https://arxiv.org/abs/1809.07763`.

Brandon Greenwell. pdp: An R package for Constructing Partial Dependence Plots. *The R Journal*, 9(1):421–436, 2017.

Jos Hernndez-Orallo. ROC curves for regression. *Pattern Recognition*, 46(12):33953411, 2013. doi: 10.1016/j.patcog.2013.06.014.

Ulf Johansson, Cecilia Snstrd, Ulf Norinder, and Henrik Bostrm. Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Medicinal Chemistry*, 3(6): 647–663, 2011. doi: 10.4155/fmc.11.23.

Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 2008. doi: 10.18637/jss.v028.i05.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, New York, NY, USA, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?"*. ACM Press, 2016. doi: 10.1145/2939672.2939778. URL `https://arxiv.org/abs/1602.04938`.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems.* MIT Press, 2015.

Agnieszka Sitko and Przemyslaw Biecek. *The Merging Path Plot: adaptive fusing of k-groups with likelihood-based model selection*, 2017. URL `https://arxiv.org/abs/1709.04412`.

Mateusz Staniak and Przemyslaw Biecek. *Explanations of Model Predictions with live and breakDown Packages*, 2018. URL `https://arxiv.org/abs/1804.01955`.